SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I6P110

Volume 11 Issue 6, 114-121, June 2024
© 2024 Seventh Sense Research Group®

Original Article

Detection of Fake and Real Violence Using Hierarchical CNN Model

Lucky Rajpoot[1], Rosy Madaan[2]

[1,2]*School of Engineering and Technology, Manav Rachna International Institute of Research and Studies, Haryana, India.*

[1]*Corresponding Author : lucky06jpn@gmail.com*

Received: 16 April 2024 Revised: 26 May 2024 Accepted: 11 June 2024 Published: 29 June 2024

Abstract *- This investigation delves into the intersection of deep learning and image processing for early detection and classification of violence, with a primary focus on differentiating between movie fights (staged or fake) and true violence. Leveraging the "Violence and Non-violence Images Dataset," along with the collected movie fight images dataset, the proposed methodology involves Training Model3 (Hierarchal combination of Model1 and Model2). The hierarchy enhances performance and significantly improves specificity scores, even in a dataset biased toward nonviolence cases. The proposed model achieves an impressive accuracy of 98.33%, showcasing its potential for crime detection.*

Keywords *- CNN, Domain Transfer Learning (DTL), Domain Data Augmentation (DDA), Deep learning, LSTM.*

1. Introduction

In current years, the over-provision of digital content and the prevalent use of image classification techniques have led to remarkable advancements in various domains. One notable application is the use of image classification in detecting violence, a critical endeavor in enhancing public safety and security. Women, in general, have benefited greatly from these.[1] However, the accuracy of such systems is not without its challenges, as instances of misclassification, particularly in the context of movie scenes depicting fights, can lead to the generation of false alerts. Movie scenes often involve staged fights that may be misclassified as real-life violence by image classification algorithms, resulting in unnecessary and potentially disruptive notifications. To address this issue, researchers have turned to advanced methodologies such as hierarchical classification combined with a transfer learning approach. This innovative approach aims to improve the precision of violence detection systems by refining the classification process and minimizing false positives, specifically in scenarios where distinguishing between staged and real violence is crucial.[2][3]

Hierarchical classification involves organizing classes into a hierarchical structure, allowing the model to learn and distinguish between different levels of features. In the context of movie fights and violence detection, a hierarchical classification system enables the algorithm to recognize the nuances between staged fights and real violence, leading to more accurate and context-aware results [3, 4]. The proposed research on the hierarchical classification of movie fights and violent images holds significant promise for various applications. Firstly, it can enhance the performance of surveillance systems by reducing false alerts enabling security personnel to focus on genuine threats. Additionally, in the realm of content moderation, online platforms can benefit from more precise filtering mechanisms, preventing the unnecessary removal of content due to misclassifications.[5]

The intersection of image classification and violence detection presents both challenges and opportunities. By leveraging hierarchical classification with a transfer learning approach, We aim to refine the accuracy of detection systems, mitigating the misclassification of movie fights and improving the overall reliability of violence detection algorithms in various applications.[6][7]

2. Literature Review

In response to the escalating demand for efficient monitoring systems in light of increasing violence cases, a research paper by Vieira et al. presents a solution employing low-cost Convolutional Neural Networks (CNNs) for the automatic recognition of suspicious events. The study utilizes a curated dataset, combining instances of violent behavior and non-violent acts across diverse environments. Notably, MobileNet-v2 emerges as the most accurate among the three mobile CNN architectures, achieving a high accuracy of 91.63%. The paper delves into the trade-off between accuracy and the total number of parameters, asserting that mobile CNNs prove effective even in interfaces with restricted processing capability. The research makes a valuable contribution to the practical challenges associated with deployment costs and processing speed on embedded systems, particularly in real-world applications of deep learning for violence recognition.[8]

In a comprehensive review, Host and Ivašić-Kos explore Human Action Recognition (HAR) research in global team sports, encompassing soccer, volleyball, hockey, basketball, table tennis, tennis, and badminton. The study notes a significant surge in research over the last four years, with soccer standing out as the most researched sport. Challenges in HAR for sports, such as simultaneous actions, cluttered backgrounds, and varied perspectives, are discussed. The review highlights commonly used feature representations, with optical flow being prevalent, and explores Machine Learning (ML) and Deep Learning (DL) methods. The authors emphasize the importance of creating specialized databases for implementing HAR methods, contributing valuable insights into existing frameworks, challenges, and trends in HAR research for various sports.[9]

Addressing the imperative need for analyzing surveillance videos in public and industrial security, Ullah, et al. focus on violence detection within the Industrial Internet of Things (IIoT) context. The proposed VD-Net framework integrates artificial intelligence into an IIoT-based system to efficiently detect violence, utilizing a lightweight CNN for initial information gathering and a ConvLSTM for detailed investigation in the cloud. VD-Net surpasses modern violence detection methods, showing a 3.9% increase in accuracy. The paper introduces a new real-world industrial surveillance dataset and conducts a comparative analysis with state-of-the-art methods, categorizing them based on learning strategies. The study positions VD-Net as an effective solution for violence detection in industrial setups, addressing challenges of computational complexity and viewpoint variations[10,11].

In the domain of video action recognition, a survey paper by Wu et al. underscores the importance of understanding human behaviors through video action recognition, particularly in sports analytics. The survey covers a wide array of sports, both team and individual, providing a comprehensive overview of video action recognition methods. The paper explores challenges posed by sports-related data, fast-paced actions, and the complexities of recognizing actions in team sports. It includes a practical aspect with the development of a toolbox using PaddlePaddle, a deep learning platform tailored to support action recognition in specific sports like football, basketball, table tennis, and figure skating. The paper establishes the groundwork for a thorough examination of video action recognition in sports analytics, offering insights into existing frameworks and challenges.[12]

In the realm of automated video surveillance systems, Himeur et al. underscore the critical role of these systems in ensuring public security during events with large crowds. Recognizing challenges inherent in Deep Learning (DL) algorithms, the paper introduces innovative solutions. DTL and DDA aim to ease training processes, enhance model generalizability, and overcome data scarcity issues. The paper provides a comprehensive overview of existing DTL- and DDA-based video surveillance methods, addressing their benefits and challenges by outlining future perspectives. It positions itself as a valuable resource to contribute critical insights into the current state of research in DTL and DDA-based video surveillance.[13]

Asad et al. introduce a novel method for detecting violent actions in videos, specifically addressing the challenge of continuous human observation in autonomous surveillance systems. The proposed approach combines spatial and temporal features extracted from sequential frames using a CNN and LSTMs. Notably, the introduction of "Wide-Dense Residual Blocks (WDRB)" effectively learns combined spatial features, showcasing high accuracy compared to modern methods. The research emphasizes the significance of transfer learning, additional residual blocks, and LSTM units in enhancing the model's capability to detect various types of violent actions in videos, contributing to the advancement of autonomous surveillance technology.[14]

Tommasi et al. present a comprehensive study on violence detection in videos, introducing the CrimeNet neural network. The core problem addressed is achieving high average precision in violence detection while minimizing false alarms. CrimeNet, based on the ViT architecture and NSL with adversarial training, significantly outperforms previous models, reducing false positives to nearly zero. The study rigorously tests CrimeNet on challenging violence-related datasets, demonstrating substantial improvements in ROC AUC across different datasets. Leveraging optical flow for pre-processing and incorporating NSL for adversarial learning contribute to CrimeNet's state-of-the-art performance. The paper concludes by recognizing CrimeNet's achievements, emphasizing the significance of NSL and adversarial learning, and suggesting avenues for future research.[15]

Finally, Ullah et al. introduce an intelligent violence detection approach tailored for industrial video surveillance. The proposed method combines lightweight CNNs, optical flow features, and LSTM networks for effective sequential pattern analysis. The paper highlights the potential effectiveness of the proposed approach in addressing the challenges of violence detection in industrial surveillance settings. The contributions to existing datasets and improvements in accuracy underscore the significance of this approach for enhancing video surveillance capabilities in industrial contexts.[16]

3. Dataset Description

The dataset, due to limitations of the available ones, was collected from diverse sources, they include online opensource datasets and images.

The first dataset is of violence and non-violence images. It was taken from an opensource platform. This dataset contains only two directories: Non-violence (which contains 5231 real-life situations images like eating, sports activity, singing, etc, and this directory does not have any violent situations) and the other directory, Violence (contains 5842 images with severe violence in various situations) the dataset was extracted from 1000 real life videos of violence and nonviolence activities like eating, walking, sports and more. The dataset captures real street fight situations, providing a realistic and challenging set of scenarios for violence recognition models. These videos encompass various environments and conditions, adding complexity to the dataset and enhancing its relevance to real-world applications.

The second dataset was a collection of movie fight scenes augmented to fit the scale. The dataset originally had only 1000 images, they were each augmented to balance the classes.

4. Methodology

4.1. Augmentation and Preprocessing

The dataset had an imbalance in the number of images for each class, so the first step we took was augmentation; we applied transformations like cropping, zooming, width shift (3 px max), mirroring, etc.

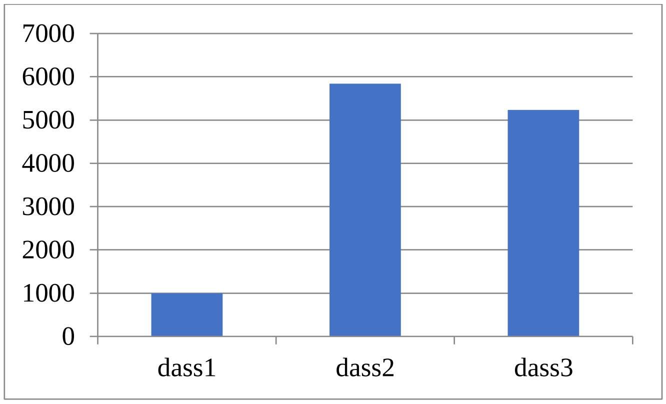

Fig. 1 Imbalance in number of classes here: Class 1 refers to the movie fight class, class 2 refers to the non-violence class, class 3 is the violence class

The preprocessing pipeline for the images entails a multi-step procedure. Initially, the images are resized to a standardized format of 300 by 300 pixels, ensuring uniformity in their dimensions. Subsequently, after a train test split, a noise reduction technique is implemented using the Gaussian blur function sourced from the OpenCV2 library, contributing to the refinement of image quality.

To further enhance the visual characteristics, a contrast adjustment stage is introduced. This involves the extraction of YUV features through the application of the OpenCV2 cvtColor function. Specifically, the luminance component (Y channel) is isolated, and the contrast is normalized using the cv2.equalizeHist operation. This strategic contrast adjustment serves to refine the overall perceptual quality of the images.

The resultant images, now endowed with standardized dimensions, reduced noise, and optimized contrast, are then seamlessly integrated into the subsequent layers of the processing pipeline, thus laying the foundation for downstream analysis or model training.

4.2. Training Model 1

The model, denominated as "Model 1", represents a CNN designed specifically for violence detection. Leveraging the ResNet18 architecture, transfer learning is employed to harness the pre-trained weights from a ResNet18 model while tailoring the final layers to address the targeted task. The input layer is configured with dimensions of 224 x 224 x 3 to accommodate the RGB color channels, and the `include_top` parameter is set to false for ResNet18.

The transfer learning process involves extracting convolutional and dense layers from the ResNet18 architecture, thereby capitalizing on learned features from a diverse dataset. Subsequently, fine-tuning is conducted on a new dataset encompassing classes denoted as "Violence" and "Non-violence." To mitigate overfitting, a custom dropout layer with a rate of 0.1 is applied.

Post-transfer learning layers, the results traverse a max-pooling layer to downsample spatial dimensions in the representation, thereby reducing computational complexity and highlighting salient features. A sigmoid activation function is then applied to the output layer, facilitating the transformation of raw network output into probabilities. Its formula is as depicted below:

$$S(x) = \frac{1}{1 + e^{-x}}$$

The sigmoid activation mainly helps with binary classification, and the model performance is more easily controlled than with multiple-class classification problems.

Throughout the training process, the Adam optimization function is identified as the most suitable optimization technique, with a learning rate set to 0.0001. Additionally, an early stopping technique is implemented for regularization purposes. Evaluation metrics such as accuracy, precision, recall, or F1-score are employed to assess the model's performance.

In essence, the transfer learning CNN architecture of Model1 endeavors to adeptly capture and classify patterns pertaining to violence and non-violence in images. It derives advantages from the knowledge accrued by the ResNet18 model during its pre-training on a broader image recognition task. The model performed poorly with the classification of violence and non-violence images. Its accuracy was 86.03%

4.3 Training Model 2

The model denominated as "Model2" represents a CNN designed specifically for violence detection. The input layer processing and the parameter taken for Model2 are the same as Model1. For training the model, the input dataset is non-violence and the movie fight and the model will classify the dataset into these two classes. The activation function of the hidden layer is the same, but a sigmoid activation function is applied to the output layer, facilitating the transformation of raw network output into probabilities.

In essence, the proposed transfer learning CNN architecture endeavours to adeptly capture and classify patterns pertaining to Movie fights and non-violence in images. It derives advantages from the knowledge accrued by the ResNet18 model during its pre-training on a broader image recognition task. The model performed a little poorly with the classification of violence and nonviolence images. Its accuracy was 99.38%

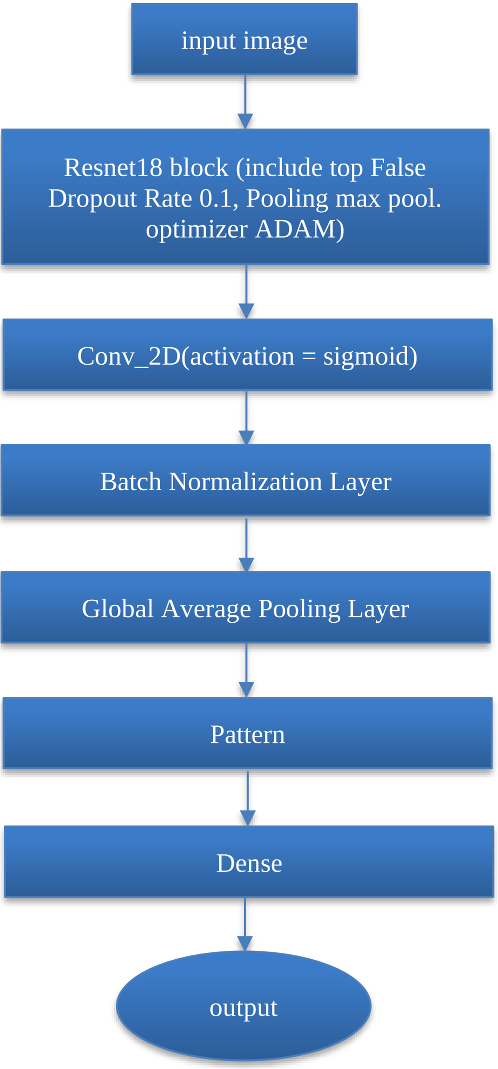
input image
Resnet18 block (include top False Dropout Rate 0.1, Pooling max pool. optimizer ADAM)
Conv_2D(activation = sigmoid)
Batch Normalization Layer
Global Average Pooling Layer
Pattern
Dense
output

Fig. 2 Model 1 training

4.4. Training Model 3

The model 3 is a hierarchal classifier ensemble of Model 1 and Model 2. Its functioning is as described below.

Model 1 undertakes the processing of input images, yielding probabilities or scores associated with the violence and movie fight classes. The outcomes generated by Model 1 serve as inputs to Model 2, which assumes the responsibility of further refining the classification. Upon classification by Model 1, if an image is designated as depicting violence, an additional layer of validation ensues via Model 2_violence. Conversely, in the event of classification as movie Fights, the outcome is directly conveyed as such. Images not conforming to either of these classifications are designated as violence-class images by default.

The decision to categorize a given probability as indicative of violence or movie fights is contingent upon a predefined threshold. This threshold acts as a discriminative criterion, determining the point at which a probability surpasses the defined threshold and is consequently assigned to either the violence or movie fights category. This architecture achieved an accuracy of 98.33%. This multi-step architecture, involving Model 1 and subsequent validation through Model 2, aims to enhance the precision and granularity of image classification, particularly in the nuanced domains of violence and movie fight identification.

4.5. Evaluation Methods

Some commonly used metrics to evaluate the performance of classification models are used to evaluate the performance of model are:

4.5.1. Accuracy (AC)

Definition: AC quantifies the inclusive correctness of the model by computing the ratio of correctly projected instances to the total number of instances.

Formula

$$Ac = \frac{Total\ Number\ of\ Predictions}{Number\ of\ Correct\ Predictions}$$

Interpretation

AC is a general measure of correctness but may not be suitable for imbalanced datasets where the classes have significantly different sizes.

4.5.2. PE

Definition: PE, also famed as ”positive predictive value”, quantifies the AC of ”positive predictions”. It is the ratio of correctly predicted positive instances to the total number of predicted positives.

Formula

$$PE = \frac{True\ Positives}{True\ Positives + \ False\ Positives}$$

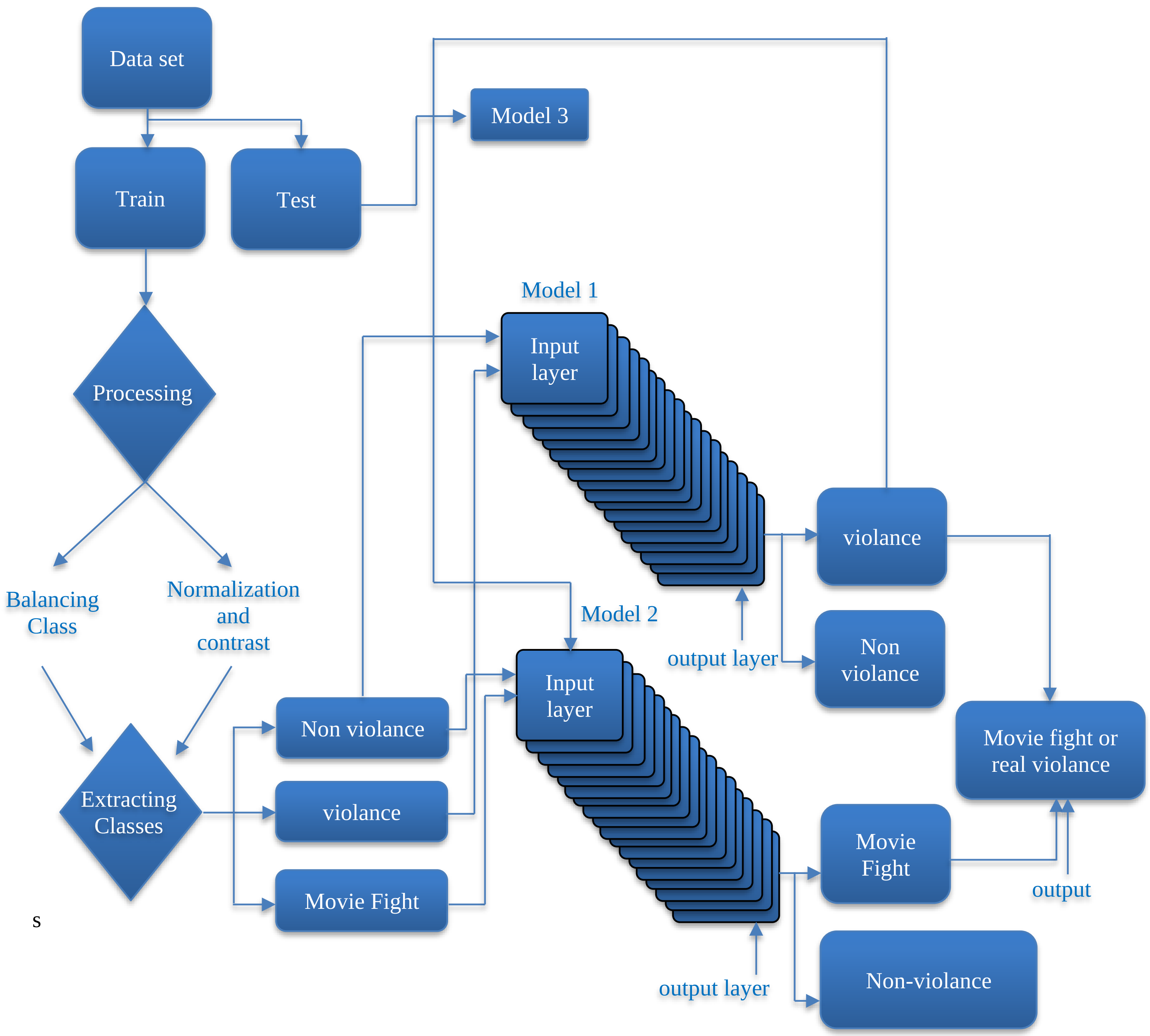

Fig. 3 Flowchart of methodology and Model3 architecture

Interpretation

PE becomes increasingly significant when the cost or number of false positives is high, and there is a need to minimize false positive predictions.

4.5.3. SEN (Sensitivity or True Positive Rate)

Definition: SEN quantifies the adeptness of the model to capture all the positive instances. It is the fraction of correctly predicted positive instances to the total actual positives.

Formula

$$\text{SEN} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Interpretation

SEN becomes increasingly significant when the cost or no. of false negatives is high, and there is a need to minimize instances of the positive class being missed.

4.5.4. F-score (FS)

Definition: The FS is the HM of PE and SEN, providing a uniform measure of both. It is especially useful when there is an uneven class distribution.

Formula

$$FS = 2 \times \frac{PE \times SEN}{PE + SEN}$$

Interpretation

The FS glues PE and SEN into a single metric, with higher values indicating a better balance between the two.The models were each evaluated with these factors. A summary of the table of results is as follows.

Table 1. AC, SEN, FS and PE of the models

	Accuracy	Precision	Sensitivity	F1-score
Model 1	86.03	84.9	87.53	84.03
Model 2	99.38	98.23	98.46	98.92
Model 3	98.33	99.05	98.42	98.76

The table 1 presents performance metrics for three models: Model 1, Model 2, and Model 3. Model 1 demonstrates an accuracy of 86.03%, precision of 84.9%, sensitivity (recall) of 87.53%, and an F1-score of 84.03%. In contrast, Model 2 exhibits superior performance with an accuracy of 99.38%, precision of 98.23%, sensitivity of 98.46%, and an impressive F1-score of 98.92%. Model 3 also performs well, achieving an accuracy of 98.33%, precision of 99.05%, sensitivity of 98.42%, and an F1-score of 98.76%. Overall, Model 2 demonstrates the highest accuracy and balanced precision, sensitivity, and F1-score, making it the top-performing model among the three.

4.6. ResNet18

ResNet18, a variant of the residual network architecture, is designed to overcome challenges associated with training extreme DNNs. Its distinctive components play crucial roles in its effectiveness.

ResNet18 introduces the innovative concept of residual blocks, each containing two convolutional layers. The inclusion of shortcut connections facilitates the learning of identity mappings, addressing the vanishing gradient problem and enabling the training of very deep networks. The layered architecture involves stacking multiple residual blocks, progressively increasing the number of filters to deepen the network.

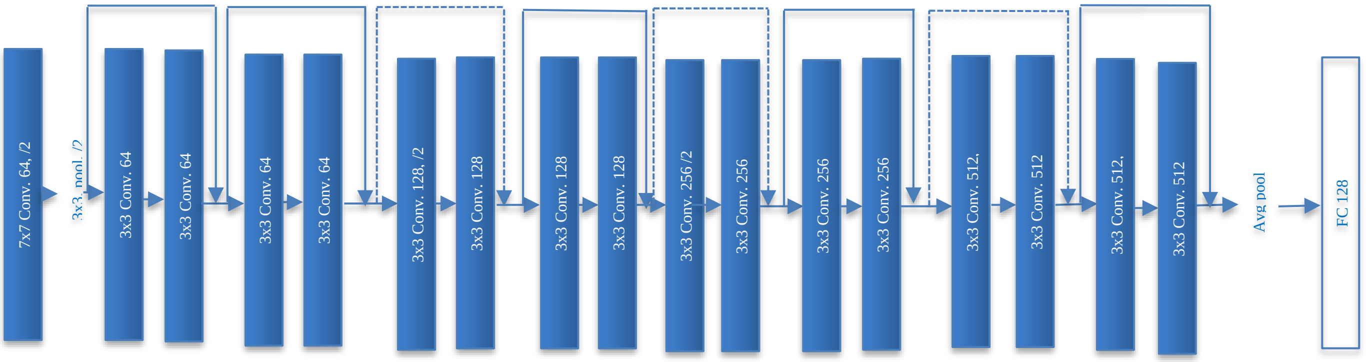

Fig. 4 Resnet18 architecture

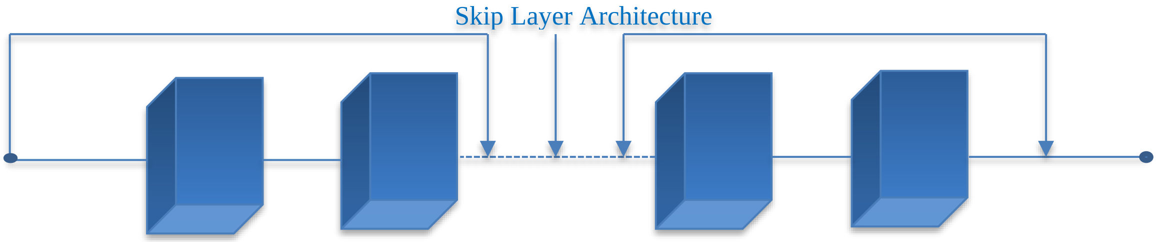

Fig. 5 Skip connections

Convolutional layers, fundamental to ResNet18, perform spatial convolutions with 3x3 filters within residual blocks. Batch normalization is employed to normalize layer inputs, stabilizing and accelerating training by mitigating internal covariate shift. The Rectified Linear Unit (ReLU) serves as the activation function, introducing non-linearity to capture complex relationships in the data.

$$\mathrm{ReLU}(x) = \max(0, x)$$

Max pooling layers downsample spatial dimensions, reducing computational complexity. Fully connected layers at the network's end perform classification based on high-level features learned by convolutional layers. Global Average Pooling (GAP) is utilized before the fully connected layers to provide a compact representation for classification.

Skip connections, or shortcut connections, facilitate gradient flow during backpropagation by directly connecting input to output, aiding in the training of deep networks. ResNet18 often serves as a pre-trained model, leveraging weights learned from large datasets like ImageNet for transfer learning on specific tasks where labeled data is limited.

5. Results and Discussion

The proposed model, denoted as Model 3, demonstrated exceptional performance within the constraints of the limited dataset, exhibiting an impressive accuracy of 99.38%. This remarkable accuracy is substantiated by specific performance metrics, including 1187 true positives, 997 true negatives, 11 false positives, and 19 false negatives. The model underwent training for a duration of 300 epochs, showcasing its resilience and ability to generalize effectively.

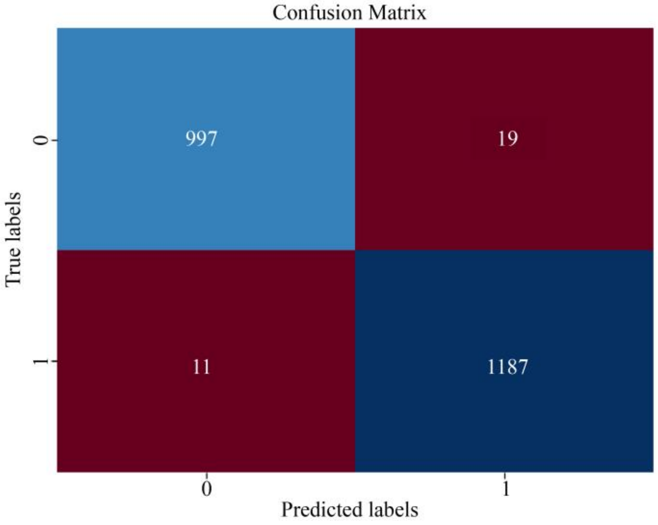

Fig. 6 Confusion matric of the predicted result

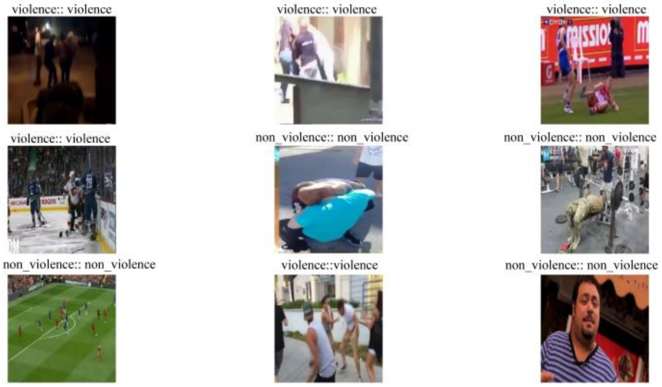
violence:: violence
violence:: violence
violence:: violence
violence:: violence
non_violence:: non_violence
non_violence:: non_violence
non_violence:: non_violence
violence::violence
non_violence:: non_violence

Fig. 7 Some predicted results

In comparison, ResNet18 models, when trained on the entire dataset, exhibited a range of accuracies between 86.43% and 92.56%. However, an intriguing observation emerged when the dataset was stratified, and the models were tailored for binary classification tasks.

In this context, the accuracy witnessed a noteworthy improvement, ranging up to 98.33%. Its precision, sensitivity and f1 score are approximately 99.05%, 98.4% and 98.7%. This disparity in performance raises intriguing questions, and several factors may contribute to this phenomenon.

Firstly, the heightened accuracy of Model 3 could be attributed to enhanced confidence in detecting instances belonging to the non-violence class. The model might have acquired a nuanced understanding of features characterizing non-violence, contributing to its heightened accuracy in discerning such instances.

Secondly, the simplicity inherent in the binary classification problem undertaken by Model 3 might have facilitated a more effective learning process. Binary classification tasks often allow models to focus on differentiating between two distinct classes, potentially leading to more refined and accurate representations.[14]

Lastly, the hierarchical nature of movie fight classification, where the outcome depends on the accuracy of both Model 3 and the complementary model responsible for the non-violence class, could further contribute to the observed increase in accuracy. The combined hierarchical approach ensures a comprehensive evaluation of the input, considering both violence and non-violence aspects.

In conclusion, the exemplary performance of Model 3 on the limited dataset underscores the nuanced interplay between dataset characteristics, model architecture, and the intricacies of the classification task. The observed accuracy improvements in binary classification scenarios illuminate the potential benefits of tailored model design in addressing specific nuances within the dataset.[5][17]

6. Future Scope

The successful performance of Model 3 on the limited dataset, as evidenced by its remarkable accuracy, precision, sensitivity, and F1 score, opens avenues for promising future research and applications. Despite the achievements realized within the constraints of a limited dataset, it is crucial to acknowledge the inherent limitations imposed by dataset size, which may impact the model's generalizability. This limitation prompts an imperative consideration for the expansion of the dataset, fostering a more comprehensive understanding of diverse instances and further enhancing the model's robustness.

One notable avenue for future exploration lies in the realm of video processing, particularly for the classification of staged fights versus authentic altercations. Model 3, with its adeptness in discerning violence and non-violence in images, could be extended to video processing scenarios. The integration of temporal information across frames could fortify the model's capacity to distinguish between staged fights, prevalent in cinematic contexts, and real instances of violence. The temporal analysis of video sequences holds the potential to capture nuanced patterns that are indicative of genuine altercations, thus contributing to a more sophisticated and context-aware classification system.[6][7]

Moreover, the proposed model could benefit from further optimization and fine-tuning. Hyperparameter tuning, such as adjusting learning rates or exploring alternative optimization algorithms, may enhance the model's convergence and generalization capabilities. Additionally, the exploration of advanced model architectures beyond ResNet18, such as deeper or more specialized networks, could be a direction for future improvement. In the context of practical applications, the deployment of such a violence detection model holds great potential for security and surveillance systems. Integrating the model into surveillance cameras or monitoring systems could provide real-time alerts or assistance in identifying potential security threats. This not only enhances the efficiency of security measures but also contributes to the overall safety and well-being of individuals within monitored environments.

Furthermore, considerations for ethical implications and bias in the dataset should be rigorously addressed. Future research should prioritize the exploration of strategies to mitigate biases in training data and model predictions, ensuring fair and unbiased outcomes across diverse demographic groups.

References

[1] Marcella Papini et al., "The Role of Deep Learning Models in the Detection of Anti-Social Behaviours towards Women in Public Transport from Surveillance Videos: A Scoping Review," *Safety*, vol. 9, no. 4, pp. 1-26, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Viktor Huszar, "*Possibilities and Challenges of the Defensive Use of Artificial Intelligence and Computer Vision Based Technologies and Applications in the Defence Sector*," Doctoral PhD Dissertation, pp. 1-21, 2023. [Google Scholar]

[3] Md Golam Morshed et al., "Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities," *Sensors*, vol. 23, no. 4, pp. 1-40, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ryo Hachiuma, Fumiaki Sato, and Taiki Sekii, "Unified Keypoint-Based Action Recognition Framework via Structured Keypoint Pooling," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, pp. 22962-22971, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Elarbi Badidi, Karima Moumane, and Firdaous El Ghazi, "Opportunities, Applications, and Challenges of Edge-AI Enabled Video Analytics in Smart Cities: A Systematic Review," *IEEE Access*, vol. 11, pp. 80543-80572, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Nadia Mumtaz et al., "An Overview of Violence Detection Techniques: Current Challenges and Future Directions," *Artificial Intelligence Review*, vol. 56, p. 4641-4666, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Muhammad Awais, and Sara Durrani, "Violence Activity Detection Classification-A Review," *International Conference on Scientific and Academic Research*, vol. 1, pp. 139-144, 2023. [Publisher Link]

[8] Joelton Cezar Vieira et al., "Low-Cost CNN for Automatic Violence Recognition on Embedded System," *IEEE Access*, vol. 10, p. 25190-25202, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Kristina Host, and Marina Ivasic-Kos, "An Overview of Human Action Recognition in Sports Based on Computer Vision," *Heliyon*, vol. 8, no. 6, pp. 1-25, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Fath U. Min Ullah et al., "AI-Assisted Edge Vision for Violence Detection in IoT-Based Industrial Surveillance Networks," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5359-5370, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Abbas Z. Kouzani, "Technological Innovations for Tackling Domestic Violence," *IEEE Access*, vol. 11, pp. 91293-91311, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Fei Wu et al., "A Survey on Video Action Recognition in Sports: Datasets, Methods and Applications," *IEEE Transactions on Multimedia*, vol. 25, pp. 7943-7966, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] Yassine Himeur et al., "Video Surveillance Using Deep Transfer Learning and Deep Domain Adaptation: Towards Better Generalization," *Engineering Applications of Artificial Intelligence*, vol. 119, pp. 1-34, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Mujtaba Asad et al., "Multi-Frame Feature-Fusion-Based Model for Violence Detection," *The Visual Computer*, vol. 37, p. 1415-1431, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Fernando J. Rendón-Segador et al., "Crimenet: Neural structured Learning Using Vision Transformer for Violence Detection," *Neural Networks*, vol. 161, pp. 318-329, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Fath U. Min Ullah et al., "An Intelligent System for Complex Violence Pattern Analysis and Detection," *International Journal of Intelligent Systems*, vol. 37, pp. 10400-10422, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Dalia Andrea Rodríguez et al., "A Systematic Review of Computer Science Solutions for Addressing Violence Against Women and Children," *IEEE Access*, vol. 9, pp. 114622-114639, 2021. [CrossRef] [Google Scholar] [Publisher Link]